

Database

OrthoClusterDB: an online platform for synteny blocks

Man-Ping Ng^{†1}, Ismael A Vergara^{†1}, Christian Frech¹, Qingkang Chen¹,
Xinghuo Zeng², Jian Pei² and Nansheng Chen^{*1}

Address: ¹Department of Molecular Biology and Biochemistry, Simon Fraser University, 8888 University Drive, Burnaby, V5A 1S6, Canada and
²School of Computing Science, Simon Fraser University, 8888 University Drive, Burnaby, V5A 1S6, Canada

Email: Man-Ping Ng - pennyng119@gmail.com; Ismael A Vergara - iav@sfu.ca; Christian Frech - cfa24@sfu.ca;

Qingkang Chen - cqk118@gmail.com; Xinghuo Zeng - xzeng.sfu@gmail.com; Jian Pei - jpei@cs.sfu.ca; Nansheng Chen^{*} - chenn@sfu.ca

^{*} Corresponding author [†]Equal contributors

Published: 23 June 2009

Received: 10 February 2009

BMC Bioinformatics 2009, **10**:192 doi:10.1186/1471-2105-10-192

Accepted: 23 June 2009

This article is available from: <http://www.biomedcentral.com/1471-2105/10/192>

© 2009 Ng et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: The recent availability of an expanding collection of genome sequences driven by technological advances has facilitated comparative genomics and in particular the identification of synteny among multiple genomes. However, the development of effective and easy-to-use methods for identifying such conserved gene clusters among multiple genomes—synteny blocks—as well as databases, which host synteny blocks from various groups of species (especially eukaryotes) and also allow users to run synteny-identification programs, lags behind.

Descriptions: OrthoClusterDB is a new online platform for the identification and visualization of synteny blocks. OrthoClusterDB consists of two key web pages: *Run OrthoCluster* and *View Synteny*. The *Run OrthoCluster* page serves as web front-end to OrthoCluster, a recently developed program for synteny block detection. *Run OrthoCluster* offers full control over the functionalities of OrthoCluster, such as specifying synteny block size, considering order and strandedness of genes within synteny blocks, including or excluding nested synteny blocks, handling one-to-many orthologous relationships, and comparing multiple genomes. In contrast, the *View Synteny* page gives access to perfect and imperfect synteny blocks precomputed for a large number of genomes, without the need for users to retrieve and format input data. Additionally, genes are cross-linked with public databases for effective browsing. For both *Run OrthoCluster* and *View Synteny*, identified synteny blocks can be browsed at the whole genome, chromosome, and individual gene level. OrthoClusterDB is freely accessible.

Conclusion: We have developed an online system for the identification and visualization of synteny blocks among multiple genomes. The system is freely available at <http://genome.sfu.ca/orthoclusterdb/>.

Background

Accumulating evidence suggests that genes within a genome are not randomly distributed. Instead, they form various types of conserved gene clusters, such as operons [1,2], genes co-regulated by common transcription mech-

anisms [3], and genes co-expressed in a same tissue type such as muscle [4]. The recent availability of an expanding collection of genome sequences driven by technological advances has facilitated genome-wide detection of these functional gene clusters through comparative genome

analysis [5]. However, the development of effective and easy-to-use methods for identifying such conserved gene clusters among multiple genomes—synteny blocks—that at the same time host databases of these synteny blocks lags behind.

The term synteny has been used to refer different concepts in the past. Initially, synteny was used to indicate that genes are located on the same chromosome [6]. Recently, synteny has been more generally used to describe conservation, and syntenic genes have been generally taken as genes co-localized within conserved genomic blocks among related genomes [7]. There are further differences regarding the level of conservation. For example, some define two genomic sequences as a synteny block as long as they contain orthologous gene sets regardless of their order [8] or the existence of insertion/deletions [9]. In this paper, we generally follow recent definitions of synteny block and define it as a "genomic region of conserved gene content". We distinguish between "perfect synteny blocks" (genomic regions of perfectly conserved gene content, including gene order and strandedness) and "imperfect synteny blocks" (genomic regions of imperfectly conserved gene content, order or strandedness).

Most methods developed in the past years for detecting synteny blocks cannot be generally applied because they fail in one or more of the following tasks: (A) comparing multiple genomes; (B) detecting synteny blocks containing interruptions (mismatches); (C) considering strandedness (orientation) of genes; and (D) handling one-to-many orthologous relationships (reviewed in [10]). To overcome the above limitations, we have recently developed a program, OrthoCluster, for synteny block detection [10].

To make it easy for users to run OrthoCluster and to interpret the output, we have now developed a web server, OrthoClusterDB <http://genome.sfu.ca/orthoclusterdb/>, which provides an easy-to-use web interface to OrthoCluster and immediate access to synteny blocks that have been precomputed with OrthoCluster. Currently, only two synteny detection methods—Cinteny [11] and SyMAP [12]—also provide servers for online access and access to databases.

Construction and content

The OrthoClusterDB website consists of the following two key web pages: *Run OrthoCluster* and *View Synteny*. The *Run OrthoCluster* web page enables users to run OrthoCluster with their own genome annotation files and correspondence files (containing orthologous relationships among all input genomes) to identify synteny blocks. The *View Synteny* web page allows users to browse through pre-computed synteny blocks between up to

three genomes at the genome, chromosome and gene level. In addition to these two pages, OrthoClusterDB also has a *Download* page, which provides users with the datasets used for generating the pre-computed results and OrthoCluster executables, and a *Help* page, which includes answers to frequently asked questions, protocols for using *Run OrthoCluster* and *View Synteny* pages, and the OrthoCluster tutorial.

Utility and discussion


Run OrthoCluster

The *Run OrthoCluster* page allows users to run OrthoCluster online using their own genome annotation and correspondence files to identify synteny blocks among a large number of genomes (Figure 1). Before running OrthoCluster online, users are recommended to check whether their genomes of interest are already included in the precomputed datasets in the *View Synteny* page. We provide details regarding where genome annotations are obtained together with release version number for each genome so that users can track down the data source for accurate analysis and comparison.

Two types of files are needed as input, the *genome file* and the *correspondence file* (both are plain tab-delimited text files). A genome file contains all genes and their coordinates in an input genome, while the correspondence file contains orthologous relationships among all input genomes. Users can define and modify parameters for running OrthoCluster, such as block size, order and strandedness of genes within synteny blocks, and inclusion/exclusion of nested synteny blocks resulting from one-to-many orthologous relationships. Notably, even though most users run OrthoCluster using genes and their orthologous relationships as input, OrthoCluster can be used to process any type of genomic feature (or genetic markers) as long as their orthologous relationships are provided.

Users can upload two or more input genomes. The first input genome ("Genome 1") is by default taken as the *reference genome* and the rest are referred to as the *target genomes*. By default, perfect synteny blocks will be generated.

The main part of the result page consists of the *Genome Painter* image that displays an overview of detected synteny blocks between a reference genome and target genome(s) at the chromosome/contig-level (Figure 2). This is achieved by first partitioning the reference genome into segments of different colors. (A) For reference genomes with 50 or less chromosomes/contigs but more than one chromosome/contig, each chromosome/contig gets assigned a different color and is shown in a separate column with its corresponding name. (B) For reference



[Release 2](#)

| | | | | |
|------|------------------|--------------|----------|------|
| Home | Run OrthoCluster | View Synteny | Download | Help |
|------|------------------|--------------|----------|------|

Sample Run: two genomes

*Job ID

[Upload Files](#)

*Genome 1

*Genome 2

Genome 3:

[Add more Genome files](#)

*Correspondence File:

Chr/Contig Order

Chr/Contig Order

Chr/Contig Order:

* Required fields

▼ [Parameters](#)

[Order and Strandedness](#)

[Synteny Block Size](#)

[Mismatches](#)

[Blocks Produced](#)

☒ -rs
 ☐ -r
 ☐ -s
 ☐ -r -s
 ☐ None

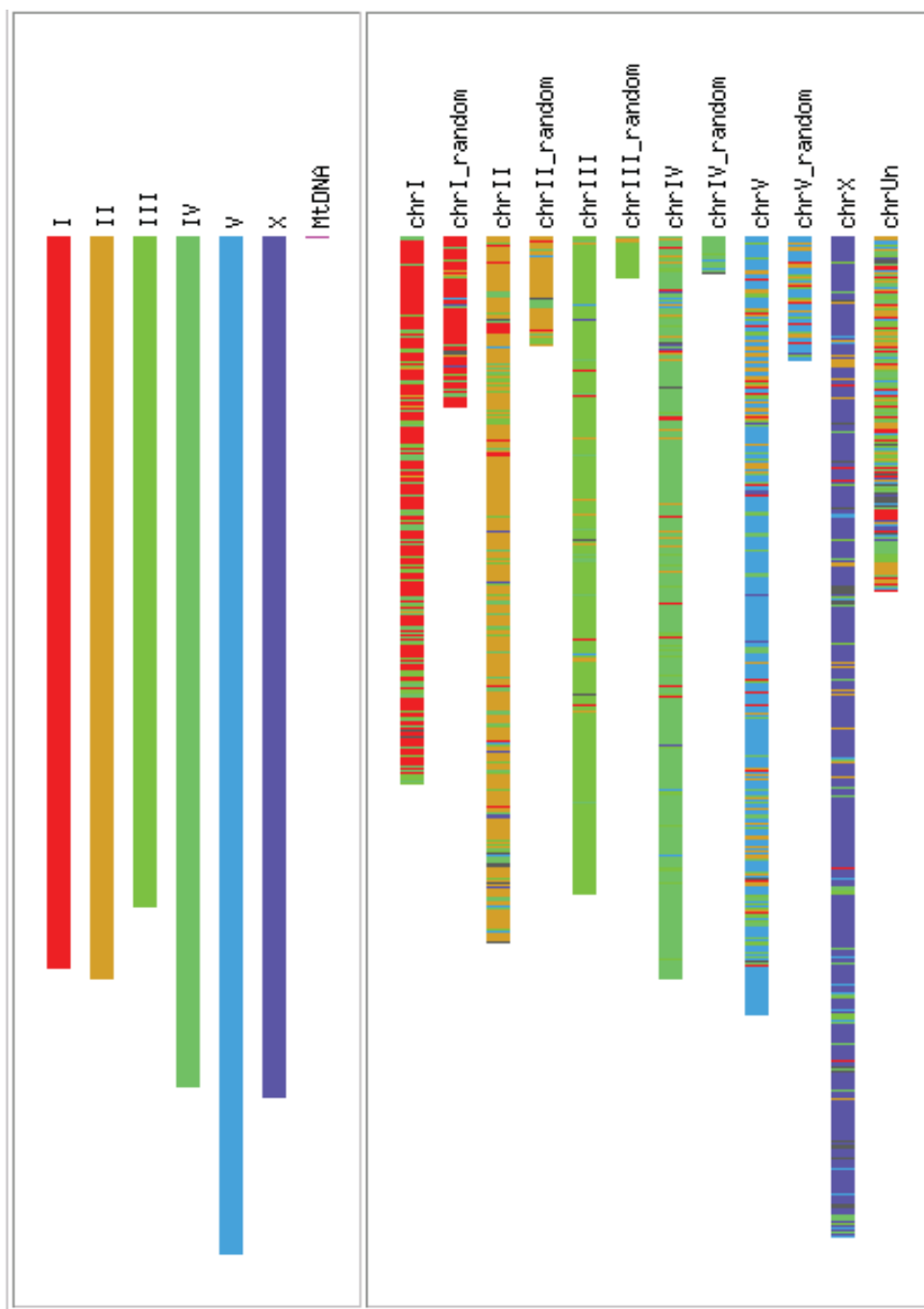
Lower Bound: Upper Bound:

-i: -ip:

-o: -op:

☒ Sort output blocks (-x)
☐ Show non-nested blocks only

Figure 1
 Web interface of the Run OrthoCluster page showing the input parameters.

**Figure 2**

A sample output *Genome Painter* image, with a link for output download and a link to *GBrowse*. In this example, *C. elegans* is the reference genome and *C. briggsae* is the target genome.

genomes with more than 50 chromosomes/contigs but fewer than 256 chromosomes/contigs, chromosomes/contigs are drawn with a continuous color gradient and without displaying their name for clarity. (C) For reference genomes containing more than 256 chromosomes/contigs, only the first 256 largest chromosomes/contigs are assigned unique colors and chromosomes/contigs beyond this number will be assigned the same color (black). (D) For reference genomes that are composed of only one chromosome/contig, the chromosome/contig is colored in a rainbow-spectrum manner. Detected synteny blocks are then highlighted within target genomes by drawing all syntenic regions in the color of their corresponding segment in the reference genome. The gray color in the target genomes indicates that no corresponding synteny blocks have been found in that region. By default, the order of the chromosomes/contigs displayed is sorted by size.

The *Genome Painter* image is particularly useful for visualizing overall conservation of different genomes. For example, as illustrated in Figure 3, there is an obvious large inversion between the *Pseudomonas aeruginosa* PAO1

genome and the genomes of *Pseudomonas aeruginosa* PA14 and *Pseudomonas aeruginosa* PA7, as reported previously [13]. In contrast, the genomes of *Pseudomonas aeruginosa* PA14 and *Pseudomonas aeruginosa* PA7 are generally very similar.

Each synteny block within the *Genome Painter* image is clickable and cross-linked to a genome browser for gene-level view of the chromosomal/contig region containing that block. We use the Generic Genome Browser (GBrowse) [14] for that purpose, a widely used genome browser program (Figure 4). Users can enter the GBrowse view either by clicking on the link in the result summary table, or by directly clicking on the color-coded synteny blocks in the genome painter image. We configured GBrowse to display the genes of the reference genome in one track and the corresponding synteny blocks of the target genomes in separate tracks. Each synteny block displayed in the genome browser in the target genome is cross-linked to another genome browser, in which the target genome is displayed as the reference for that synteny block.

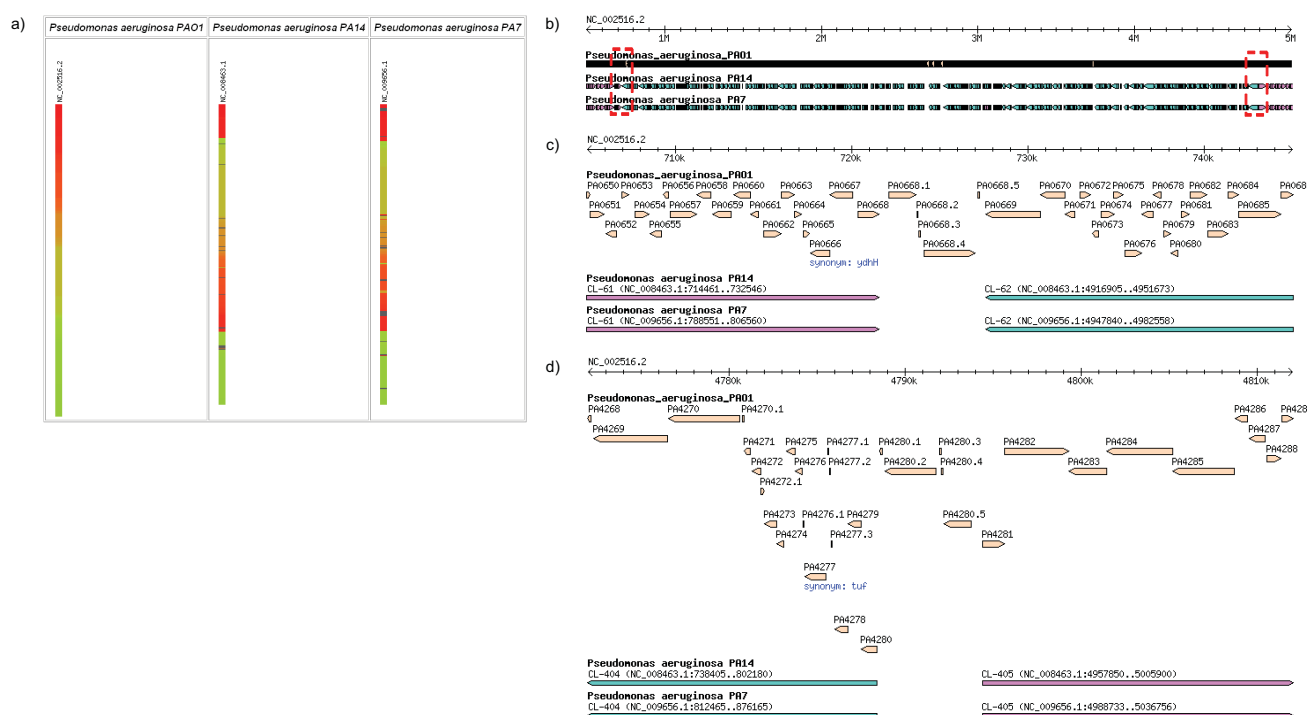


Figure 3

Large inversion in *Pseudomonas aeruginosa* genomes. a) Genome Painter image of three *P. aeruginosa* genomes, showing a large inversion of the two target genomes with respect to the reference genome (*P. aeruginosa* PAO1). b) GBrowse image of the large inverted region. The two junctions are surrounded by red dashed boxes. c) GBrowse image of the left-most junction of the inverted region. d) GBrowse image of the right-most junction of the inverted region.

C. elegans synteny

Showing 21.89 kbp from III, positions 1,558,872 to 1,580,757

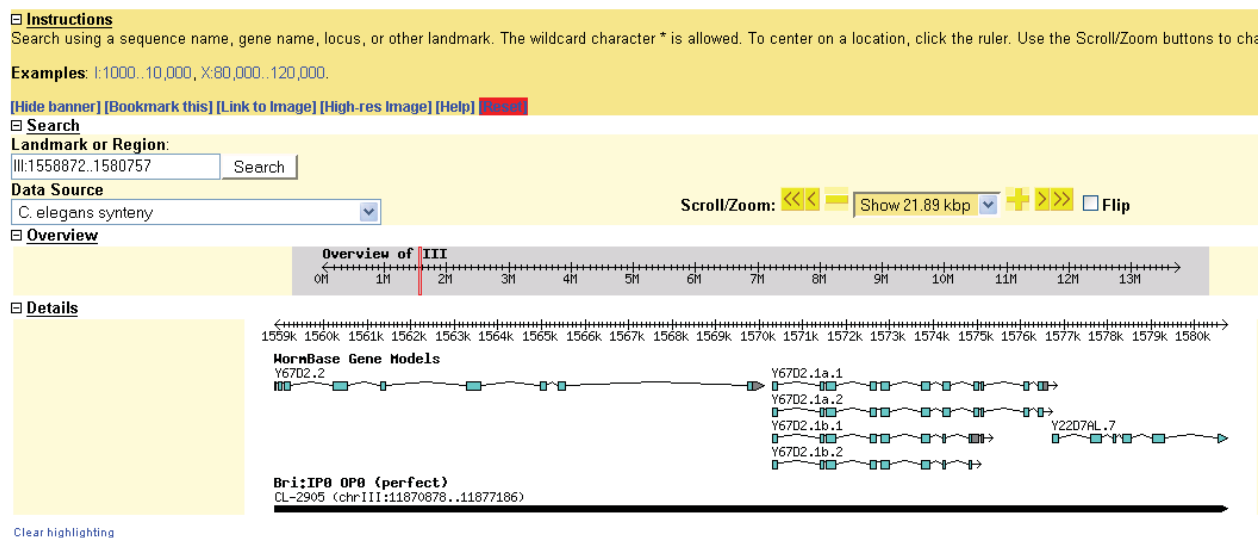


Figure 4

GBrowse-based synteny browser with *C. elegans* as reference genome and *C. briggsae* as target genome. The first track shows the WormBase gene model for *C. elegans*, and the second track is the syntenic block detected in *C. briggsae*. CL-2905 is the syntenic block ID assigned by OrthoCluster, and the number in brackets next to the ID refers to the chromosome location of the block in *C. briggsae*.

The *Run OrthoCluster* page also allows users to redefine the default behavior of OrthoCluster as well as the *Genome Painter* output. First, users can modify the order of chromosomes/contigs in the *Genome Painter* image by uploading a simple text file containing all the chromosome/contig names in a desired order. Format details can be found in the *Help* page. Second, users can generate various types of imperfect synteny blocks by varying the parameters of OrthoCluster, such as the minimum and maximum number of genes within the block, number/percentage of in-map mismatches (i.e. genes with known but non-syntenic orthologs) and out-map mismatches (genes without known orthologs) in a block, and preservation of the relative order and strandedness of orthologous genes within each synteny block. Additionally, the user is allowed to display non-nested synteny blocks only. Nested synteny blocks within larger blocks occur because of one-to-many orthologous relationships, which are usually present in the correspondence file and which OrthoCluster considers simultaneously at the moment of generating the synteny block.

View synteny

Genomes of some species are of general interest. To facilitate identification of synteny blocks between these

genomes, we have created the *View Synteny* page, where users can select predefined genomes of interest for synteny identification and examination. In the current release (Release 2), five groups of genomes are available (Figure 5): *Pseudomonas* (14 genomes), *Plasmodium* (6 genomes), *Caenorhabditis* (2 genomes), *Drosophila* (12 genomes), and Mammals (20 genomes). For these groups, genome files were preloaded and formatted on the web server. The correspondence files for running OrthoCluster are generated on the fly on the web server by parsing precomputed InParanoid [15] results (for two genomes) or by running MultiParanoid [16] (for multiple genomes).

To start a job in the *View Synteny* page, users first select a group of interest and then a reference genome within this group. Once a reference genome is selected, up to two target genomes can be chosen from the same group. Users may choose to identify perfect (no mismatches allowed) or imperfect (containing 5% in-map mismatches, 10% out-map mismatches) synteny blocks. The result page format is the same as that of the *Run OrthoCluster* page.

There are three major differences between the *View Synteny* page and the *Run OrthoCluster* page. First, in the *View Synteny* page, users do not need to prepare input files,

ORTHOCLUSTERDB

[Release 2](#)

Home Run OrthoCluster **View Synteny** Download Help

Select Group

Select Reference Genome

Select Target Genome(s)
(Up to 2 selections)

Select Conservation Option
☒ Perfect (-ip 0 -op 0)
☐ Imperfect (-ip 5 -op 10)

Figure 5
Web interface of View Synteny showing currently available groups of genomes for selection.

making it easier for the user to get quick results for the species of interest. Second, in the *View Synteny* page, genes in the genome browser are linked to their corresponding gene pages in public databases, such as WormBase <http://www.wormbase.org/> [17] for the *Caenorhabditis* group, FlyBase <http://www.flybase.org/> [18] for the *Drosophila* group, or Ensembl <http://www.ensembl.org/> [19] for the Mammals group. This makes it easy for following up individual genes within synteny blocks in more detail. Third, results from all jobs submitted via the *View Synteny* page are stored permanently in the database (MySQL) of the web server so that results will be immediately returned the next time users try to identify synteny blocks among the same genomes.

Download

The *Download* page makes available the datasets used by OrthoClusterDB to generate the precomputed results, including the genome annotation files and the pair-wise correspondence files. Genome annotation files were generated based on the GFF (General Feature Format) files obtained from the corresponding model organism databases or, in case of the *Pseudomonas* genomes, from NCBI. Pair-wise correspondence files were generated using Inparanoid [15] with default settings. Also, the standalone version of OrthoCluster for Linux, MacOS, and

Windows platforms can be downloaded, allowing users to run OrthoCluster locally on their own computers.

Computational Platform

OrthoClusterDB is currently supported by a Dual Quad Core Xeon machine that has 8G RAM. The processing time for jobs submitted via *Run OrthoCluster* and *View Synteny* page depends on the number of genes contained in input genomes and the number of orthologous relationships defined in the correspondence file. For pair-wise analysis, jobs usually finish within seconds. For multiple-genome analysis with large correspondence files, jobs may take longer. Such jobs usually take up to one minute to finish on the first run. On the second run, jobs finish immediately because previous results are cached on the server. For larger jobs, users are encouraged to download OrthoCluster from the download page and run it locally.

Conclusion

Accurate and effective identification of synteny blocks provided by OrthoClusterDB will facilitate many comparative genomics analyses, including the identification of functional gene clusters, ortholog assignment, gene model improvement, identification of lineage-specific genome family expansion and contraction, as well as the characterization of various types of genome rearrange-

ment events such as insertions/deletions, inversions, transpositions, and reciprocal translocations [10]. Currently, OrthoClusterDB allows fast access to precomputed synteny blocks for 54 different genomes within 5 groups of species of general interest. Ultimately, OrthoClusterDB will be expanded to include synteny blocks for all sequenced and annotated genomes.

Availability and requirements

OrthoClusterDB is free for public access. The engine behind OrthoClusterDB, OrthoCluster, is also free and can be downloaded from the OrthoClusterDB website <http://genome.sfu.ca/cgi-bin/orthoclusterdb/download>. OrthoCluster is an effective program and can be run on a single processor. Executables are available for MacOS, PC and Linux. Source code is available upon request.

Authors' contributions

NC and JP conceived the project. MPN, IAV, CF, QC, XZ implemented the programs. MPN, IAV and NC wrote the manuscript, with input from other co-authors. All authors read and approved the final manuscript.

Acknowledgements

This work is supported by a discovery grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada to NC. MPN and QC are supported by NSERC Undergraduate Student Research Awards (USRA). NC is also a Michael Smith Foundation for Health Research Scholar.

References

- Blumenthal T, Evans D, Link CD, Guffanti A, Lawson D, Thierry-Mieg J, Thierry-Mieg D, Chiu WL, Duke K, Kiraly M: **A global analysis of *Caenorhabditis elegans* operons.** *Nature* 2002:851-853.
- Jacob F: **Operon: a group of genes with the expression coordinated by an operator.** *C R Hebd Seances Acad Sci* 1960, **250**:1727-1729.
- Johnnidis J, Venanzi E, Taxman D, JP T, CO B, DJ M: **Chromosomal clustering of genes controlled by the aire transcription factor.** *Proc Natl Acad Sci USA* 2005, **102**(20):7233-7238.
- Roy J, Stuart J, Lund J, Stuart K: **Chromosomal clustering of muscle-expressed genes in *C. elegans*.** *Nature* 2002, **418**:975-979.
- Hardison RC: **Comparative genomics.** *PLoS biology* 2003, **1**(2):E58.
- Passarge E, Horsthemke B, Farber RA: **Incorrect use of the term synteny.** *Nature genetics* 1999, **23**(4):387.
- Pevzner P, Tesler G: **Genome rearrangements in mammalian evolution: lessons from human and mouse genomes.** *Genome Res* 2003, **13**(1):37-45.
- Gregory SG, Sekhon M, Schein J, Zhao S, Osoegawa K, Scott CE, Evans RS, Burridge PW, Cox TV, Fox CA, et al.: **A physical map of the mouse genome.** *Nature* 2002, **418**(6899):743-750.
- Margulies EH, Birney E: **Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes.** *Nat Rev Genet* 2008, **9**(4):303-313.
- Zeng X, Pei J, Vergara IA, Nesbitt M, Wang K, Chen N: **OrthoCluster: a new tool for mining synteny blocks and applications in comparative genomics.** *11th International Conference on Extending Technology (EDBT), March 25-30, 2008: 2008; Nantes, France* 2008.
- Sinha AU, Meller J: **Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms.** *BMC Bioinformatics* 2007, **8**:82.
- Soderlund C, Nelson W, Shoemaker A, Paterson A: **SyMAP: A system for discovering and viewing syntenic regions of FPC maps.** *Genome Res* 2006, **16**(9):1159-1168.
- Stover CK, Pham XQ, Erwin AL, Mizoguchi SD, Warrenner P, Hickey MJ, Brinkman FS, Hufnagle WO, Kowalik DJ, Lagrou M, et al.: **Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen.** *Nature* 2000, **406**(6799):959-964.
- Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, et al.: **The generic genome browser: a building block for a model organism system database.** *Genome Res* 2002, **12**(10):1599-1610.
- Remm M, Storm CE, Sonnhammer EL: **Automatic clustering of orthologs and in-paralogs from pairwise species comparisons.** *Journal of molecular biology* 2001, **314**(5):1041-1052.
- Alexeyenko A, Tamas I, Liu G, Sonnhammer EL: **Automatic clustering of orthologs and inparalogs shared by multiple proteomes.** *Bioinformatics (Oxford, England)* 2006, **22**(14):e9-15.
- Chen N, Harris TW, Antoshechkin I, Bastiani C, Bieri T, Blasiar D, Bradnam K, Canaran P, Chan J, Chen CK, et al.: **WormBase: a comprehensive data resource for *Caenorhabditis* biology and genomics.** *Nucleic acids research* 2005:D383-389.
- Crosby MA, Goodman JL, Strelets VB, Zhang P, Gelbart WM: **FlyBase: genomes by the dozen.** *Nucleic acids research* 2007:D486-491.
- Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, et al.: **Ensembl 2008.** *Nucleic acids research* 2008:D707-714.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

